

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平4-318684

(43) 公開日 平成4年(1992)11月10日

(51) Int.Cl.⁶

G 0 6 F 15/70
7/24

識別記号

4 6 5 Z 9071-5L
8323-5B

庁内整理番号

F I

技術表示箇所

審査請求 未請求 請求項の数1(全 6 頁)

(21) 出願番号 特願平3-85556

(22) 出願日 平成3年(1991)4月17日

(71) 出願人 000006208

三菱重工業株式会社
東京都千代田区丸の内二丁目5番1号

(72) 発明者 川瀬 直人

兵庫県神戸市兵庫区和田崎町一丁目1番1号
三菱重工業株式会社神戸造船所内

(72) 発明者 川田 かよ子

兵庫県神戸市兵庫区和田崎町一丁目1番1号
三菱重工業株式会社神戸造船所内

(72) 発明者 宮本 一正

兵庫県神戸市兵庫区和田崎町一丁目1番1号
三菱重工業株式会社神戸造船所内

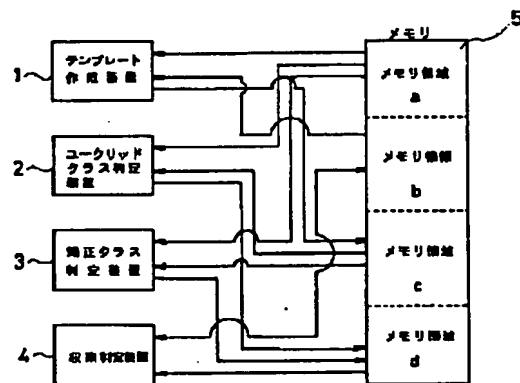
(74) 代理人 弁理士 鈴江 武彦

(54) 【発明の名称】 データ分類装置

(57) 【要約】

【目的】 データ分類装置において、データの分布状況を考慮してより正しいデータの分類を可能にする。

【構成】 テンプレート作成装置1、ユークリッドクラス判定装置2、矯正クラス判定装置3、収束判定装置4を有すると共に、特徴ベクトルデータと、どのデータがどのクラスに属するかを示すクラスデータと、処理上一時的に発生するデータを記憶するためのメモリ5を有する。分類処理の最初のステップでは種データからのテンプレートのみを利用して、ユークリッド距離でデータを一時的にクラス分けし、その後のステップでは、一時的にクラス分けしたデータからテンプレートとテンプレート補助データを生成し、それを利用して全体を矯正距離により再分類し、このステップを収束するまで繰り返すことにより、種データを基にデータ全体の分類を可能にしたものである。



Best Available Copy

【特許請求の範囲】

【請求項1】 複数の特徴量ベクトルで表現されるデータ群で、それぞれデータに対して属するクラスがある場合に、各クラスについて正しいクラスが判っている種データと、正しいクラスが判らない多数の分類データが与えられた時に、特徴ベクトルデータと、どのデータがどのクラスに属するかを示すクラスデータと、処理上一次的に発生するデータを記憶するためのメモリを有し、上記種データを基にデータ全体を分類するデータ分類装置において、クラスデータを参照して、その対応する特徴ベクトルからクラス毎にテンプレート（平均ベクトル）と、平均に関する共分散行列の固有値と固有ベクトルの全部または1部を用いたテンプレート補助データを作成するテンプレート作成装置と、クラスを識別すべきデータの特徴量ベクトルと各クラスのテンプレートとのユークリッド距離からデータのクラスを判定するユークリッドクラス判定装置と、クラスを識別すべきデータの特徴量ベクトルと各クラスのテンプレートとのユークリッド距離を、上記テンプレート作成装置で生成したテンプレート補助データを用いて、クラスの分布状態に対する相対距離に矯正し、この強制距離からデータのクラスを判定する矯正クラス判定装置と、全てのデータに対する今回のクラス判定結果が、前回のクラス判定結果と一致している時、分類が収束したとみなす収束判定装置とを具備し、最初のステップでは種データからのテンプレートのみを利用して、ユークリッド距離でデータを一時的にクラス分けし、その後のステップでは、一時的にクラス分けしたデータからテンプレートとテンプレート補助データを生成し、それを利用して全体を矯正距離により再分類し、このステップを収束するまで繰り返すことにより、種データを基にデータ全体の分類を行なうことを特徴とするデータ分類装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 本発明は、文字データ等のデータ収集に際して、オペレータが収集データ毎にデータの属するクラスを入力する手間を省くことを可能にするデータ分類装置に関する。

【0002】

【従来の技術】 少数の正解データを参考にデータ群全てを分類することを、一般的にクラスタリングと呼ぶが、そのアルゴリズムは以下になる。

【0003】 (1) r 個のクラスが存在するとして、 $\omega_1, \omega_2, \dots, \omega_r$ をそれぞれのクラスに属するデータの集合とする。初期状態においては、最初に分かっている正解データのみが ω_1 の要素である。

(2) 集合 ω_1 の要素の中心（すなわちテンプレート） m_1 を求める。全てのクラスに対して m_i を求める。

(3) 全てのデータ x について、テンプレート m_i から距離 D_i^2 が最小になる i を求める。 x を新たな分類結

果の集合 ω_i に属するものとする。

(4) $\omega_i = \omega_1$ がすべてのクラス i について成立すれば、このアルゴリズムは終了する。そうでなければ ω_1 を新たに ω_i とし(2)に戻る。従来のクラスタリングでは、距離 D_i^2 の定義は通常ユークリッド距離、すなわち、

$$D_i^2 = (x - m_i) \cdot (x - m_i)$$

を使用していた。

【0004】

【発明が解決しようとする課題】 距離としてユークリッド距離を用いる方式では、以下のようなケースにおいて誤分類する確率が高くなる。

(a) 1つのクラスが複数の集団に分かれて分布する。(図3(i))

(b) クラスの分布が偏向性を持つ。

(図3(ii))

(c) クラスによって分布のばらつき度が大きく異なる。(図3(iii)) そこで分布のばらつき度を考慮した距離を用いる必要がある。そのような距離の代表例としてはマハラノビス距離がある。その定義は次のようになる。

$$D_i^2 = (x - m_i) \cdot V_i^{-1} \cdot (x - m_i)$$

V_i : クラス i のデータの共分散行列

なお、距離 D_i としてマハラノビス距離を用いる場合には、テンプレート計算時にテンプレート補助データとして V_i を求める必要がある。しかし、次のような場合には、共分散行列 V_i が正則でない、あるいは計算上正則とはいえないためマハラノビス距離が定義できない。

(d) 特徴量ベクトルにおいて、その要素となる特徴量間の相関が非常に大きい時。(例、一方向に伸縮する幾何図形データの周長と面積。)

(e) 常に一定値をとるような無意味な特徴量がある時。(例、ひらがな2値画像データにおける周辺ピクセル値。通常0になる。)

【0005】 本発明は上記実情に鑑みてなされたもので、データの分布状況を考慮してより正しいデータの分類を行なうことができ、データ収集に際して、オペレータが収集データ毎にデータの属するクラスを入力する手間を省くことができるデータ分類装置を提供することを目的とする。

【0006】

【課題を解決するための手段】 本発明に係るデータ分類装置は、クラスデータを参照して、その対応する特徴ベクトルからクラス毎にテンプレート（平均ベクトル）と、平均に関する共分散行列の固有値と固有ベクトルの全部または1部を用いたテンプレート補助データを作成することを特徴とするテンプレート作成装置と、クラスを識別すべきデータの特徴量ベクトルと各クラスのテンプレートとのユークリッド距離から、データのクラスを判定するユークリッドクラス判定装置と、クラスを識別

3

すべきデータの特徴量ベクトルと各クラスのテンプレートとのユークリッド距離を、上記テンプレート作成装置で生成したテンプレート補助データを用いて、クラスの分布状態に対する相対距離に矯正し、この矯正距離からデータのクラスを判定する矯正クラス判定装置と、全てのデータに対する今回のクラス判定結果が、前回のクラス判定結果と一致している時、分類が収束したとみなす収束判定装置とを具備したことを特徴とするものである。

【0007】

【作用】メモリは、特徴量ベクトルを記憶する領域a、クラスデータを記憶する領域b、テンプレートとテンプレート補助データを記憶する領域c、一時的に分類されたクラス（一時的なクラスデータ）を記憶する領域dから成り立つものとする。

【0008】また、初期状態において分類データの特徴量ベクトル、種データの特徴量ベクトルが領域aに、種データのクラスデータが領域bにセットされているものとする。各装置は以下の順で動作する。まず初期動作として、

【0009】(1) テンプレート作成装置がクラスデータ（クラスID）を領域bから、またそれ（クラスID）に対応する種データ特徴量ベクトルを領域aから読み込み、テンプレートを作成し領域cに記憶する。

【0010】(2) ユークリッドクラス判定装置が、特徴量ベクトル（分類データ、種データ）を領域aから、テンプレートを領域cから読み込み、各特徴量ベクトル毎に各クラスへのユークリッド距離を計算し各特徴量ベクトルの属するクラスを判定し、クラスデータをデータIDと共に領域dに記憶し、これを領域bへコピーする。これにより領域bに記載されているクラスデータは、分類データ、種データの区別のない最新のクラスデータとなる。以下繰り返し動作として、

【0011】(3) テンプレート作成装置がクラスデータ（クラスID）を領域b（上記（2）でコピーされたもの）から、またクラスデータ（クラスID）に対応す

4

る特徴量ベクトルを全て領域aから読み込み、テンプレートとテンプレート補助データを作成し、それを領域cに記憶する。

【0012】(4) 矯正クラス判定装置が特徴量ベクトルを全て領域aから、テンプレート、テンプレート補助データを領域cから読み込み、各特徴量ベクトル毎に各クラスへの相対距離を計算し各特徴量ベクトルの属するクラスを判定し、クラスデータをデータIDと共に領域dに記憶する。

10 (5) 収束判定装置が、領域b内のクラスデータ（上記（2）でコピーされたもの）と、今回の分類結果であるクラスデータを領域dから読み込み比較する。

【0013】両方のクラスデータ（各データIDについての）が全て一致していれば収束したと判断する。このときの領域bの内容が最終的な分類結果となる。一致していない時には、今回の分類結果（dに記憶されているクラスデータ）を領域bへコピーするとともに、上記（3）から再動作させる。

20 【0014】矯正クラス判定装置において言及している相対距離とは、クラスリングにおける距離 D_1 のことであるが、問題点に対処するために次に説明する距離を用いる。あるクラスのm行m列共分散行列Vの固有値を大きい順に

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$$

とし、対応する正規化された固有ベクトルを

$$v_1, v_2, \dots, v_m$$

とする。課題で述べた理由により共分散行列が非正則、あるいは非正則に近い場合には $\lambda_i \geq \lambda_c$ ならば

$$\lambda_i \approx 0$$

30 【0015】となる整数jが存在する。そこで、 $\lambda_i > 0$ である一定の閾値 λ_c を定め、 $\lambda_i \geq \lambda_c$ となる λ_i に対応する固有ベクトルが張る部分空間に対してはマハラノビス距離 d_1 を適用し、

【0016】

【数1】

$$d_m^2 = \sum_{i=1}^m \frac{(x - m, v_i)^2}{\lambda_i} \quad (\text{但し } \lambda_m \geq \lambda_c)$$

残りの直交空間に対して、ユークリッド距離 d_2

$$d_2^2 = \sum_{i=k+1}^n \frac{(x - m, v_i)^2}{\lambda_c}$$

を導入し、これらを統合して新たな距離を定義し、

それを D_1 として用いる。

$$d_1^2 = \sum_{i=1}^m \frac{(x - m, v_i)^2}{\lambda_c} - \sum_{i=1}^m \left(1 - \frac{\lambda_c}{\lambda_i}\right) \frac{(x - m, v_i)^2}{\lambda_c}$$

5

【0017】このような距離を用いるために、テンプレート補助データとして、各クラスに対する閾値入、より大きい複数個の固有値とそれに対応する固有ベクトルを計算する。この装置の動作原理を概念的に説明すると次のようになる。

【0018】まず、初期分類はユークリッド距離を用いて行なわれるが、この分類により大半のデータは正しく分類される。そのため、その結果を用いて作成されたテンプレートとその補助データは、完全ではないがおおよそのクラスに関する分布情報を持つことになる。

【0019】次にそのテンプレートと補助データを用いた相対距離による分類は、分布情報の利用により前回よりも正しく分類される。このため、その結果を用いて作成されたテンプレートとその補助データは、前回よりも更に正確な分布情報を示していることになる。以上の動作を繰り返すことにより、データが正しく分類される比率が次第に高くなっていく。

【0020】

【実施例】以下、図面を参照して本発明の一実施例を説明する。具体例として、文字の画像データから得られた特徴量により、画像の示す文字（＝クラス）を分類するデータ分類装置を考える。

$$m^k : 1/n(k) \sum_{i=1}^{n(k)} x^k_i \quad \text{クラス } k \text{ の特徴ベクトルの平均}$$

$$V^k : 1/n(k) \sum_{i=1}^{n(k)} (x^k_i - m^k)^t (x^k_i - m^k)$$

クラス k の m^k に関する共分散行列

λ^j_i : V^k の固有値 ($\lambda^1_i \geq \lambda^2_i \geq \dots$)
 v^j_i : λ^j_i に対応する固有ベクトル ($j=1 \sim m$, m は特徴量数)

λ_i : 固有値しきい値

以下、図2のフローチャートに示す手順で動作する。

(a) 初期設定 (ステップA1)

【0023】初期状態では、分類データ (正しいクラスの判らない多数のデータ)、種データ (各クラスについて正しいクラスが判っているデータ) に対する特徴量ベクトルをメモリ5のメモリ領域aに記憶させる。データの識別のために各特徴量ベクトルには一意な番号即ちデータIDが割り振られている。ここでは分類データ、種データ合わせて1から順に番号を付ける。

【0024】メモリ5のメモリ領域bとメモリ領域dは共に同じ構成であり、クラスデータが記憶される。各クラスデータにはデータIDが伴っており、クラスデータのデータIDを参照して、対応する特徴量ベクトルを領域aから必ず見つけ出せるようになっている。初期状態では、種データに対するクラスデータのみをメモリ領域bに記憶させる。

(b) 初期テンプレートの作成 (ステップA2)

6

* 【0021】図1は本発明の一実施例に係わるデータ分類装置の全体構成を示すものである。同図において、1はテンプレート作成装置、2はユークリッドクラス判定装置、3は矯正クラス判定装置、4は収束判定装置である。また、5はメモリで、特徴量ベクトルを記憶する領域a、クラスデータを記憶する領域b、テンプレートとテンプレート補助データを記憶する領域c、一時的に分類されたクラス (一時的なクラスデータ) を記憶する領域dからなり、上記テンプレート作成装置1、ユークリッドクラス判定装置2、矯正クラス判定装置3、収束判定装置4によりアクセスされる。また、上記メモリ5には、初期状態において、分類データの特徴量ベクトル及び種データ (各クラスについて正しいクラスが判っているデータで1個以上あれば良い) の特徴量ベクトルが領域aに、種データのクラスデータが領域bにセットされているものとする。次に上記実施例におけるデータ分類方式について説明する。まず、変数の定義を行なう。

$n(k)$: クラス k のサンプルデータ数

x^k_i : クラス k のサンプルデータの特徴量ベクトル ($i=1 \sim n(k)$)

【0022】

【数2】

【0025】テンプレート作成装置1がメモリ領域bからクラスデータ (クラスID) を読み込み、そのクラスデータに対応する種データの特徴量ベクトルをメモリ領域aから読み込む。この段階ではメモリ領域bには種データに関するクラスデータのみが記憶されており、テンプレート m^k は種データから作成され、このテンプレートをメモリ領域cに記憶させる。

(c) 初期分類 (ステップA3)

ユークリッドクラス判定装置2が各データに対するクラスを判定し、初期分類を行なう。

【0026】ユークリッドクラス判定装置2は、メモリ領域aから特徴量ベクトル (分類データ、種データ) を読み込み、その各特徴量ベクトルのテンプレートをメモリ領域cから読み込み、各特徴量ベクトル毎に各クラスのユークリッド距離を求める。このユークリッド距離が最も小さくなるクラスを特徴量ベクトルに対するクラスであると判定し、その判定結果、即ち、クラスデータをデータIDと共にメモリ領域dに記憶し、これをメモリ領域bにコピーする。これによりメモリ領域bに記憶されているデータは、分類データ、種データの区別のない最新のクラスデータとなる。以上で全てのデータが一時

的にクラス分けされた状態になっている。次に以下に示すステップA4～A6の動作を繰り返して実行する。

(d) テンプレートの作成 (ステップA4)

テンプレート作成装置1がテンプレートを作成する。

【0027】テンプレート作成装置1は、メモリ領域bからクラスデータ(クラスID)(ステップA3でコピーされたもの)を読み込み、そのクラスデータに対応する種データの特徴量ベクトルをメモリ領域aから読み込む。このデータを使って各クラス毎に、平均 m^i 、固有値 λ^i 、固有ベクトル v^i を求める。平均 m^i をテンプレート、 λ^i より大きい固有値とそれに対応する固有ベクトルをテンプレート補助データとしてメモリ領域cに記憶する。

(e) データ分類 (ステップA5)

矯正クラス判定装置3が各データに対するクラスを判定する。

【0028】矯正クラス判定装置3は、メモリ領域aから特徴量ベクトルを読み込み、メモリ領域cから各特徴量ベクトルに対するテンプレート及び補助テンプレートデータを読み込み、各特徴量ベクトル毎に各クラスへの相対距離を求める。そして、相対距離が最も小さくなるクラスをその特徴量ベクトルに対するクラスであると判定し、クラスデータをデータIDと共にメモリ領域dに記憶する。

(f) 収束判定 (ステップA6)

収束判定装置4が、収束しているかどうかを判定する。

【0029】収束判定装置4は、メモリ領域b内のクラスデータ(ステップA3でコピーされたもの)と、メモリ領域dに記憶されている今回の分類結果であるクラスデータとを比較し、収束しているかどうかを調べる。メモリ領域dのすべてのクラスデータに対して、メモリ領域b内に対応するクラスデータのクラスが一致していれば収束していると判断する。このときのメモリ領域bの内容が最終的な分類結果となる。一致していない、つまり、収束していなければ、メモリ領域dに記憶している

今回の分類結果をメモリ領域bにコピーし、ステップA4から再実行する。

【0030】上記のように分類処理の最初のステップでは種データからのテンプレートのみを利用して、ユークリッド距離でデータを一時的にクラス分けし、その後のステップでは、一時的にクラス分けしたデータからテンプレートとテンプレート補助データを生成し、それを利用して全体を矯正距離により再分類し、このステップを収束するまで繰り返すことにより、種データを基にデータ全体の分類を可能にするもので、これによりデータの分類状況を考慮してより正しいデータの分類を行なうことができる。なお、本発明は、少数の正解データを参考にして、データ群全てを分類することが必要な全ての装置において適用し得るものである。

【0031】

【発明の効果】以上詳記したように本発明によれば、複数個の特徴量で表現されるデータ群で、それぞれのデータに対して属すべきクラスがある場合に、各クラスについて正しいクラスが分かっている少数個のデータと、正しいクラスが分からない多数のデータが与えられた時に、少数の正解データを元にデータ全体の分類を可能にする装置において、データの分布状況を考慮してより正しいデータの分類を行なうことができ、文字データ等のデータ収集に際して、オペレータが収集データ毎にデータの属するクラスを入力する手間を省くことができる。

【図面の簡単な説明】

【図1】本発明の一実施例に係るデータ分類装置の全体構成を示すブロック図。

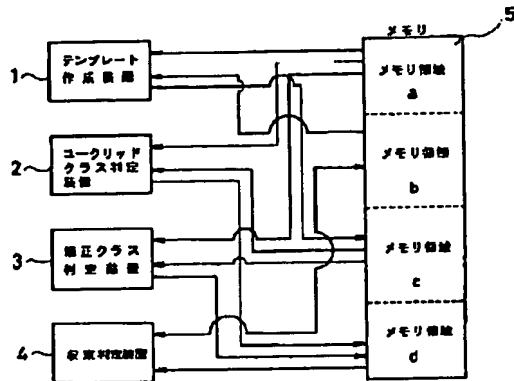
【図2】本発明の動作を示すフローチャート。

【図3】誤分類が発生し易いケースの説明図。

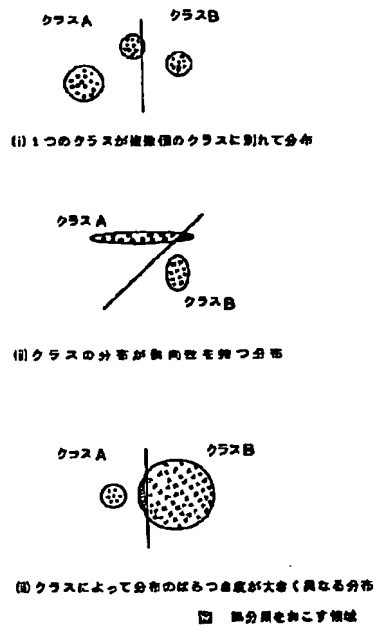
【符号の説明】

1…テンプレート作成装置、2…ユークリッドクラス判定装置、3…矯正クラス判定装置、4…収束判定装置、5…メモリ。

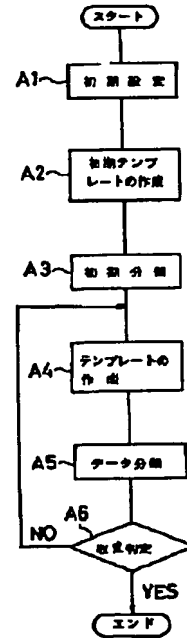
【図1】



【図3】



【図2】



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☒ FADED TEXT OR DRAWING
- ☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.